# PDF Extraction API Evaluation Template

**Team: B8**

**Team members: Sathvik, Rutuja and Sakshi**

**Link to your analysis:**

**Summary:**

We chose ConvertAPI as our enterprise PDF extractor

## 1. General Information

|  |  |
| --- | --- |
| Attribute | Details |
| API Name | ConvertAPI |
| Vendor | ConvertAPI |
| Version/Release Date | Unspecified |
| Pricing Model | Free(Limited Time)/Pay-as-you-go/Subscription |
| Licensing and Compliance | ISO 27001 Information Security Management certification |

## 2. Technical Capabilities

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | ConvertAPI | PymuPDF | Other Vendor | Other Vendor |
| File Format Support (PDF, DOCX, etc.) | PDF, DOCX, PPT, XLS, PNG, etc. | PDF, EPUB, XPS, CBR, CBZ |  |  |
| OCR (Optical Character Recognition) | High accuracy for text extraction | Not natively supported |  |  |
| Table Extraction | Yes, accurate table extraction | Supports tables but requires manual handling |  |  |
| Form Extraction | |  | | --- | |  |  |  | | --- | | Supports structured forms | | Limited support for structured forms |  |  |
| Complex Layout Support | Handles complex layouts | Basic extraction with some limitations on complex layouts |  |  |
| Multi-language Support | Supports multiple languages | Supports extraction in multiple languages |  |  |
| Scalability and Performance | Highly scalable, fast | Moderate performance, scales depending on the system's power |  |  |
| API Integration and Usability | Easy API integration with comprehensive SDKs and documentation. | No API, but easy integration with Python scripts |  |  |
| Customization Options | Customizable extraction rules | Limited customization, more manual intervention |  |  |
| Accuracy and Error Handling | High accuracy, strong error handling | Requires more manual checks for complex documents |  |  |

## 3. Business and Strategic Considerations

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Evaluation Metric | ConvertAPI | PymuPDF | Other Vendor | Other Vendor |
| Cost Efficiency (Pricing vs. Features) | Cost-effective with multiple pricing options. (Ranges from $30 - $300 per month per developer) | Free, open-source solution |  |  |
| Vendor Reputation and Stability | Less known but long serving vendor | Widely used in academic and developer communities |  |  |
| Customer Support and SLA | Strong support, clear SLA | Limited support (community-driven) |  |  |
| Security and Privacy | ISO 27001 certified, GDPR compliant | Open-source, no built-in data privacy features |  |  |
| Documentation and Training Resources | Comprehensive resources available | Extensive documentation but less formal training |  |  |
| Community and Ecosystem | Strong user community, many third-party integrations | Active community with many Python-based resources |  |  |
| Roadmap and Innovation | Large variety of functionalities and innovation in document processing | Stable with fewer new features or innovations |  |  |
| Vendor Lock-in Risk | Moderate, dependent on subscription | None, open-source and fully customizable |  |  |

## 4. Performance Metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Metric | ConvertAPI | PyMUPDF | Other Vendor | Other Vendor |
| Latency | Low, fast response times | |  | | --- | |  |  |  | | --- | | Depends on the system's performance | |  |  |
| Throughput | High throughput | Moderate throughput, dependent on hardware |  |  |
| Error Rate | Low error rates with good handling | Higher error rates for complex documents |  |  |
| Data Loss/Integrity | Minimal data loss | Possible data loss in complex layouts |  |  |

## 5. Value-Add Features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | ConvertAPI | PyMuPDF | Other Vendor | Other Vendor |
| Advanced AI/ML Capabilities | Some built-in AI features for better extraction | Lacks AI/ML capabilities |  |  |
| Pre-built Templates for Specific Use Cases | Industry-specific templates available | No pre-built templates (community support) |  |  |
| Document Classification/Tagging | Available | |  | | --- | |  |  |  | | --- | | Not supported | |  |  |
| Metadata Extraction | |  | | --- | |  |  |  | | --- | | Supports metadata extraction | | Supports metadata extraction but with limited automation |  |  |

## 6. Overall Evaluation

|  |  |  |
| --- | --- | --- |
| Attribute | Rating | Comments |
| Technical Fit | 9/10 | ConvertAPI is strong in features; PyMuPDF is solid but requires more manual intervention. |
| Business Fit | 7/10 | |  | | --- | |  |  |  | | --- | | ConvertAPI is a better choice for enterprise use; PyMuPDF works well for smaller projects. | |
| Total Cost of Ownership | 6/10 | |  | | --- | |  |  |  | | --- | | ConvertAPI can get costly over time; PyMuPDF is free but requires more setup time. | |
| Ease of Implementation and Use | 8/10 | ConvertAPI is easier to implement; PyMuPDF requires more technical expertise. |
| Vendor Reliability and Support | 8/10 | ConvertAPI has strong vendor support, while PyMuPDF relies on community support. |

## 7. Recommendations

|  |  |
| --- | --- |
| Recommendation | Details |
| Best Fit for the Use Case | ConvertAPI offers several APIs for different use cases and can be easily implemented. Its outputs are well-formatted and easily handled. PyMuPDF is more useful for small projects, but not suited for large-scale enterprise applications. |
| Further Considerations | Consider combining both APIs: ConvertAPI for complex, large-scale needs and PyMuPDF for simpler extractions with manual oversight. Also, utilize other ConvertAPI api endpoints for easier task handling. |